Session: News from ISOLand

1. Helen Conover
2. Hook Hua
3. Ted Habermann

IsoLineage Metadata at AMSR-E SIPS – Helen Conover – GHRC DAAC/AMSR-E SIPS, University of Alabama in Huntsville

* Terms for talk
  + dataset (ISO) = date file (individual science data file)
  + Product = series (ISO) (collection of data files
* AMSR-E (Advanced Microwave Scanning Radiometer – Earth Observing System
* SIPS (Science Investigator Lead operating system)
* GHRC – does provenance (how did you get this, where did it come from, how can it be used – used to be called processing history) and add metadata and QC to data
* Products = brightness temperatures, ocean products, monthly and daily ocean grids, sea ice concentration, snow depth, sea ice drift (typical NASA microwave suite)
* Capture the contextual knowledge
  + Some is already there
  + Recently – putting metadata into ISO lineage metadata model
    - Lineage so it can be added to full suite of ISO data
* Legacy data system (HDF-EOS2)
* Capture – which data products go into which for the different data projects (ex. Rain has rain and brightness temperatures)
  + SIPS provides control script – does not include science
* ISO is complex (comprehensive) – need to make friends in community
  + Only look at lineage
* Lineage Model
  + Lineage – descript source, and processing (which does down to algorithm)
  + LE = 19115-2 – they are an extension of the original model (LI) – to facilitate more detailed description of lineage
  + DQ\_DataQuality 🡪 LI\_Lineage (quality of this product – i.e. what went into making it)
* XML and ISO are verbose way to ‘saying things’ – intending to attach it to the data file (increase size)
  + Ended up with HDF SE attribute (with HDF-EOS) – this is an HDF4
* Lineage Granularity
  + Lineage info is the same across all product – capture info for a unique file (when and where it was processed)
  + At product or series level – capture attribute information
  + Keep all lineage data in each file (2 elements in each file)
  + XML “dataset” and “series”
* Lineage Model
  + Where put information – first big job (then how to say)
  + ProcessStep – high level processing description
  + Algorithm – science algorithm name, version, author, description (high level info and pointers to real data)
  + Delivered algorithm package might change but not always change the science algorithm
  + DOI and specific descriptions in Source files
  + These are done once per version of data
  + Then automated process for each data file for processing date/time/ location/ input and output files
* Question – Echo data (cloud cover) – easy map from ECHO to ISO
  + ECHO attribute or PSA – is that mapped in LE\_Algorithm?
  + Value would be somewhere else
  + Ted – can have any number of processing steps (0..\*) – can have separate ProcessStep or Algorithm
* Q – what level of granularity
  + Tried to capture the science algorithms
  + Ex. Sea ice – one processing executable, gridding, and 2 algorithm, and then snow depth
* Q – each file has many attributes
  + In provenance system – capture the attributes – map the variables to each algorithm
  + Not to the level of equation names (some have actual names and others are descriptions)
  + Did not do mapping of variable to algorithm in ISO
* Versioning
  + ESDT – doesn’t change often
  + DAP (Delivered Algorithm Package)
  + Trying to tie the processing algorithm version to the metadata
    - Includes what algorithm does, description, and author info
* DOIs
  + NASA trying to figure out how to handle DOI at ISO level
  + NASS ES difference between GCMD DIF and netCDF CF
  + Decided to combine url and DOI and then text associate is the “doi”
* Q – can use anything not use DOI
  + Yes – hence put DOI (but description is not part of Identifier)
* Use codeSpace to indicate NASA ESDIS as publisher
  + What this to be part of the NASA flavor of ISO
  + They are the authority for DOI
* Challenges – complicated, it is evolving, schemas have not been promptly provided, need to reach community consensus
* Need – NASA flavored schema, concrete examples, representations in other languages, communication (?online forum)
* Q (Aleksandar) – where get processing lineage before put into ISO
  + Red on screen = online form, talk to producer, fill in form and store in database
  + Blue = processed in house (file read events, software evocation) – parsed into database and then XML
* Q (Jennifer) – do you have a cheat sheet of summary
  + Lots info the NOAA GEO-IDE wiki
  + A lot of details – emailed works – then sorted through
  + Need online resource (don’t have fully validated XML) – will be there in a month
  + NASA is also developing their own wiki

A Practical Application Using ISO Metadata – Incorporating ISO Metadata into SMAP Data Products – Barry Weiss Hook Hua, Vance Haemmerle (JPL)

* SMAP – first NASA decadal mission
* Soil moisture
* 15 products L1 to L4 (parsed radar telemetry to carbon net ecosystem exchange)
  + Trying to create ISO for different data products (large undertaking)
* Level 1 requirement for ISO metadata (required from the top)
  + Using ISO because it is international, common representative (contextual model and encoding of it)
  + Include tools, use cases
* ISO basic concepts
  + Granule metadata = dataset
  + Collection metadata = series metadata
  + Codelist = enumerated list of accepted values
  + Profile = community agreement of particular elements
  + Extension = explicit modification (NASA is a flavor)
  + EX = Extent, LE = lineage, CI = citation
* ISO geographic standards
  + Using ISO 19130 – imagery sensor model
  + Usually talking about 19139 encoding
* UML
  + Progress Code – completed, historic obsoleted
  + SMAP needs extension points – need to mark products as “beta, stage 1, stage 4” – what to augment in code list
  + Do we need to standardize code list
    - 2 camps (1) 19139 XML (2) HDF5 group
    - Kept both
* From the Earth Science Data Model (ESDM) – all in HDF5 metadata group
  + Create crosswalk between HDF5 and ISO groups
  + Ex. Lineage would be subgroup in HDF5
  + Lineage group include attitude, ephemeris, antenna pointing, …
  + Renamed MI\_Identifier as identified\_product\_doi for DOI
* Started with UML diagrams from ISO and expanded where needed for SMAP
  + What were the extensions that were needed
  + Then generated spreadsheet – provided mapping between ISO to HDF to ESDM
  + ESDM defines gaps, ISO only beginning and ending
  + Ex. Extent – needed to add a vertical extent
* Spreadsheet provided more than exercise – did the mapping programmatically
  + Mapping used by converter to generate extra files for crosswalk
* For series metadata - Delivered by data architect
* For dataset – problem = automation
  + Spreadsheets are the first step
  + Use info to automatically inject the correct fields into the ISO from HDF
  + Able to reduce dependencies to only HDF5 libraries – simplified things
* Saxon used for crosswalk (transforms needed for each flavor)
  + Decoupling science software and metadata dialect
  + That means if the dialect changes
* Q (Peter) - Has program to move from HDF to ISO
  + Not writing software – writing rules (saxon)
  + XSL – it is an open source tool (apache product)
  + Kept as simple as possible
  + Ted – also have transform from OpenDAP land to ISO and NC-ISO – translates from one XML to another
* Q - (Peter) – rules defines the fields? - yes
* Q – what about binary data
  + HDF group – h5dump – ignores all data arrays (only dumps out metadata)
* DOI and UID
  + MD\_Identifier has been updated to be a formal class (changed) – identify if DOI
* SMAP extension
  + Additional attributes (ex. Run time parameters)
    - Eos and echo additional attributes
    - Issue – couples the type and the values – need to repeat type definition throughout (sometimes doesn’t want to repeat)
  + Only one citation for algorithm
* Validation
  + XML data binding tools
  + Ted suggested schematron approach – use rules (popular in ISO community)
    - Ex. Width needs to be followed by high
* Limitations in HD5 (1.8 library)
  + In hierarchy – group names have to be unique (can’t represent arrays of groups)
    - But arrays are common in ISO
  + H5dump – UDT (user defined data types)
    - Not fully supported – become text blogs
* NASA flavor recommendations
  + Acquisition information – some belong to granule and some series
  + Namespaces
* ISO is cutting edge… to NASA
* Lessons
  + Easing into ISO
  + ISO deeply nested
  + Simplicity – ex. Only HDF5 (easier to use Matlab)
  + Need flavor
* Q – how benefiting from international flavor
  + 19000 series – cover geographic but not mission specific
  + Flavor is a community agreement (not changing standard) – use same extension – these options
* Q (Erin) – how different than a profile
  + Call it what you want (NASA likes flavor)
  + Flavor – is a code list (Erin)
* Instrument, platform, processing – ISO revised every 5 years – implement now as extension and then add to discussion for future (community process of extensions)
* Q – will the standards have evolved in time for SMAP mission (currently using previous version)
  + Ted – ISO has standard mechanism to extend itself
* Q (Alek) – how much larger
  + It is in the noise (10-70 k) –
  + Helen had 100 k files that only had 10k spot

Wikis, Rubrics, Views and Connections: An Integrated Approach to Improving Documentation – Ted Habermann, Anna Milan – NOAA/NESDIS/NGDC

* Tools are on top of web accessible folders
  + Also use portal (external view)
* Here help people who are creating metadata to improve it to better understand connections
* NOAA wiki – NOAA EDM (old GEO IDE) <http://geo-ide.noaa.gov/wiki>
* Wiki
  + Discussion pages – include examples – have explanations – first things created on wiki
  + ISO explorer – for class/element – structure/order/ alternatives – help people editing metadata
  + Pages were created based on community input (based on questions to Ted)
  + Training – approach to learning ISO – building blocks (structured paths through wiki content) – wiki more like encyclopedia… Ted uses them like books
* Wiki Navigation
  + Categories – important – automatic to group pages (many-many & sub-categories)
    - Work like a home page
  + ISO Explorer – has classes of FGDC (things need to be in the right order – not the same as the UML)
* Many of the pages are updated mainly by Anna and Ted, but other people too… it is an ongoing effort
* Web Accessible Folder
  + Folders available from website
  + People mange metadata in databases
  + Web access folder are then like a cache – people can harvest
  + Titles (with stars related to score), Links, Sources, last update, views (get data, FAQ, HTML, fields, comments, KML)
* HTML view – able to link to wiki from each of these views
* Metadata evaluation – rubric
  + Mechanism for evaluation – here completeness of metadata
  + 1) use attribute convention for data discovery (ACDD)
  + 2) defined by Ted’s group
  + Rubric made of spirals made of fields… linked to wiki – dynamic user guide
    - Red = bad, green = good – other information provided via urls (best practice)… … opportunities for improvement
  + Each record has score… this is an evaluation tool
* Connections – community has lots of dialects (or metadata standards)
  + ESIP wiki – documentation connections
  + How to document difference connects (ex. People – provide different dialects xpaths) – if you know more > talk to Ted
* Q (Hook) – is this a NOAA manage/operated or community
  + Ted controls who can contribute
* Q – we want to control/understand what document is being referred to in metadata – references in documents may include URLs – do you see a way to control obsolete data in a rubric
  + maintenance of links in metadata record
  + tools sit on web folders that check links
  + also – prefer xlink and then links controlled elsewhere
  + use something similar to link checking websites – work with series
  + recommend not using link in granule
  + Use resolvers (doi:)
* Q - can the rubric provide guidance
  + Guidance but not control
* Q – DOI landing page
  + When someone resolves DOI it goes to that page – can be created in metadata
* Q – can landing page provide permanent link
  + Can easily extract links in a file and put them elsewhere – if permanent it can be a permanent landing page
  + Need to be actively manage/testing these